

STRUCTURED DOCUMENT RETRIEVING DEVICE

Publication number: JP10198697

Publication date: 1998-07-31

Inventor: NAKATSUYAMA HISASHI

Applicant: FUJI XEROX CO LTD

Classification:

- international: G06F17/21; G06F17/27; G06F17/30; G06F17/21; G06F17/27;
G06F17/30; (IPC1-7): G06F17/30; G06F17/21; G06F17/27

- European:

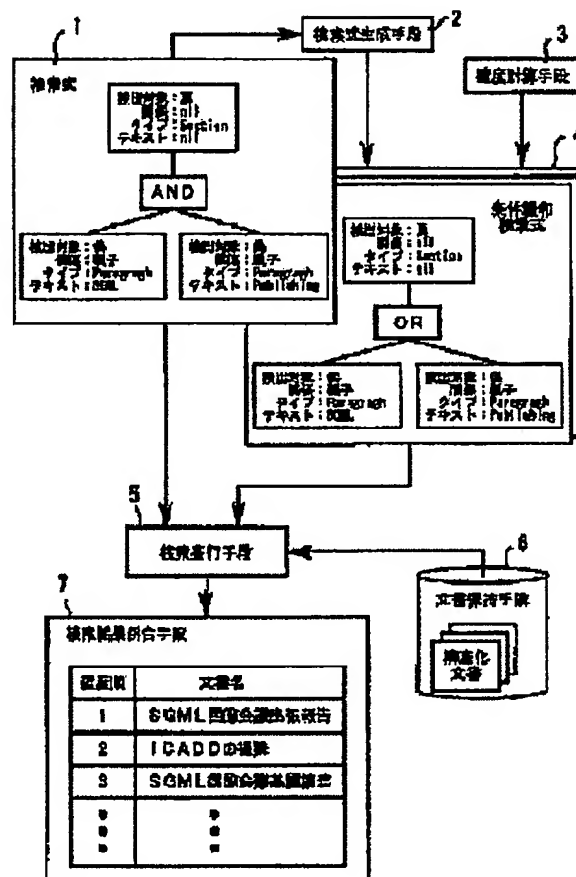
Application number: JP19970004269 19970114

Priority number(s): JP19970004269 19970114

Report a data error here

Abstract of JP10198697

PROBLEM TO BE SOLVED: To make it possible to retrieve documents at high reproducibility even in the case of retrieving plural documents having respectively different logical structures. **SOLUTION:** When a retrieval expression 1 is inputted, a retrieval expression generating means 2 rewrites retrieving conditions indicated in the expression to stepwisely moderate conditions and generates a condition eased retrieval expression 4. An accuracy calculating means 3 calculates accuracy indicating the accuracy of a retrieval result based on the expression 4 in accordance with the rewritten condition executed for the formation of each expression 4. A retrieval execution means 5 retrieves structured documents stored in a document storing means 6 based on the inputted retrieval expression 1 and the retrieval expression 4 generated by the means 2. A retrieved result merging means 7 merges retrieved results by arranging them in order from the result having the highest accuracy. Consequently a document of which logical structure is not correctly prepared can also be retrieved and the reproducibility of documents can be improved.



Data supplied from the esp@cenet database - Worldwide

(51)Int.Cl.⁸

識別記号

F I

G 0 6 F 17/30
17/27
17/21

G 0 6 F 15/403 3 3 0 B
15/20 5 5 0 E
5 7 0 N
15/40 3 7 0 A

審査請求 未請求 請求項の数6 O L (全 14 頁)

(21)出願番号

特願平9-4269

(22)出願日

平成9年(1997) 1月14日

(71)出願人 000005496

富士ゼロックス株式会社

東京都港区赤坂二丁目17番22号

(72)発明者 中津山 恒

神奈川県足柄上郡中井町境430 グリーン

テクなかい 富士ゼロックス株式会社内

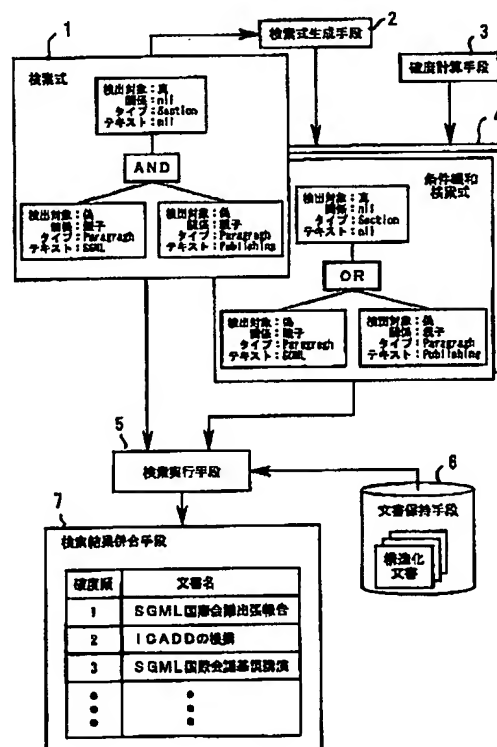
(74)代理人 弁理士 服部 毅巖

(54)【発明の名称】 構造化文書検索装置

(57)【要約】

【課題】 論理構造の異なる複数の文書に対する検索においても、高い再現率で検索可能にする。

【解決手段】 検索式1が入力されると、検索式生成手段2が、検索式に示された検索条件を段階的に緩やかな条件に書き換え、条件緩和検索式4を生成する。確度計算手段3は、各条件緩和検索式4を生成するのに行った書き換えの内容に応じて、条件緩和検索式4による検索結果の確からしさを示す確度を計算する。検索実行手段5は、入力された検索式1と検索式生成手段2により生成された条件緩和検索式4のそれぞれにより、文書保持手段6内の構造化文書を対象として検索を実行する。検索結果併合手段7は、検索実行手段5による検索結果を、確度の高い順に並べて併合する。これにより、正しく論理構造が作成されていない文書も検索することができ、再現率が向上する。



【特許請求の範囲】

【請求項1】 構造化文書を検索対象とする構造化文書検索装置において、

文書構造中のノードのタイプ、ノードの内容、ノードの属性、ノード間の構造上の関係によって記述された検索式が入力されると、前記検索式に示された検索条件を段階的に緩やかな条件に書き換えた条件緩和検索式を生成する検索式生成手段と、

前記条件緩和検索式を生成するのに行った書き換えの内容に応じて、前記条件緩和検索式による検索結果の確からしさを示す確度を計算する確度計算手段と、

入力された前記検索式及び前記条件緩和検索式に基づいて検索を行う検索実行手段と、

前記検索実行手段による検索結果を、確度の高い順に並べて併合する検索結果併合手段と、

を有することを特徴とする構造化文書検索装置。

【請求項2】 前記検索式生成手段は、予め定められた限界値よりも高い確度の条件緩和検索式のみを生成することを特徴とする請求項1の構造化文書検索装置。

【請求項3】 書き換え規則ごとに基準確度を割り当てる確度割当手段をさらに有し、

前記確度計算手段は、前記確度割当手段が割り当てた基準確度に基づいて、前記条件緩和検索式の確度を計算することを特徴とする請求項1記載の文書検索装置。

【請求項4】 構造化文書を検索対象とする構造化文書検索装置において、

文書構造中のノードのタイプ、ノードの内容、ノードの属性、ノード間の構造上の関係によって記述された検索式が入力されると、前記検索式内の各検索条件を個別の部分式とし、部分式毎にそれぞれの条件を満たした文書部品の取り出しを行う部分式評価手段と、

前記部分式評価手段で取り出された文書部品が、どのような部分式の条件を満たしているかに基づいて、各文書部品の確からしさの度合いを示す確度を計算する確度計算手段と、

確度の高い文書部品の順に検索結果を出力する検索結果出力手段と、

を有することを特徴とする文書検索装置。

【請求項5】 前記検索結果出力手段は、予め定められた限界値よりも高い確度の文書部品のみを検索結果とすることを特徴とする請求項4記載の構造化文書検索装置。

【請求項6】 部分式の種類ごとに基準確度を割り当てる確度割当手段をさらに有し、

前記確度計算手段は、前記確度割当手段が割り当てた基準確度に基づいて、各文書部品の確度を計算することを特徴とする請求項4記載の構造化文書検索装置。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】 本発明は構造化文書を対象と

した文書検索を行う構造化文書検索装置に関し、特に複数の文書型から生成された文書を検索対象とする構造化文書検索装置に関する。

【0002】

【従来の技術】 構造化文書では、文書の内容は論理構造と呼ばれ、章、節、図などの複数の文書構成要素からなる木構造で表現される。図19は論理構造の例を示す図である。このような論理構造61はまったく自由に作成してよいのではなく、文書型と呼ばれる構文規則に沿って作成される。

【0003】 図20は文書型の例を示す図である。この文書型60の中で、矩形のノードは要素の型（要素型）を定義している。ノードのラベルは、要素型の名前を示している。同一の名前をもつノードの実体は同一の要素型である。したがって、図20の「節」という名前の要素型は、再帰的に定義されていることになる。

【0004】 楕円で示したノードは要素のつながりを定義する。このノードを構築子と呼ぶ。SEQノードは、それにつながるノードのインスタンスがその順に生成されることを示している。REPノードは、それにつながるノードのインスタンスが1回以上生成されることを示す。OPTノードは、それにつながるノードのインスタンスが、出現してもしなくてもよいことを示す。CHOノードは、それにつながるいずれか1つのノードのインスタンスが生成されることを示す。ここで「インスタンス」とは、この文書型に基づき生成される文書の要素を示す。

【0005】 図20の文書型の定義を書き下すと、次のようになる。「記事」は1つ以上の「節」からなり、「節」は「見出し」と0個以上の「段落」または「図」および0個以上の「節」からなる。前述のように、「節」は入れ子になってよい。図19の論理構造61は、図20の文書型60の制約を満たしている。

【0006】 図20では簡単な文書型の例を示したが、実用規模の文書型は大規模であり、要素型の数が数百に及ぶことも珍しくない。文書型は、データベースで言えばスキーマに相当する。即ち、文書の要素の意味と、要素間の関係とを記述したものが文書型である。データベースの処理がスキーマにしたがって行われるのと同様に、構造化文書の処理は文書型の情報に基づいて行われる。例えば、文書型にしたがって割り付け指示を定義しておき、文書インスタンスと割り付け指示とを入力として、文書割り付けが行われる。もう1つの例として、既存の文書群から必要な部分を適宜抽出、それらを合成して新たな文書を作成する例があげられる。このとき、必要であれば、新規な部分を入れ込むこともある。このような処理においては、必要な部分を特定する検索処理と、新たに構成した文書が所望の形態であるかを検査する検証処理などに文書型の情報が用いられる。

【0007】 文書型に基づいて構造化文書を作成するに

は、論理構造を直接ユーザに提示するネイティブエディタを用いることができる。ネイティブエディタを利用するには、構造化文書そのものと、文書作成に用いる文書型に精通している必要がある。

【0008】ユーザが構造化文書や文書型に精通していない場合、テキストエディタや、印刷イメージとほぼ同じ画面表示を行う文書作成ソフトウェア(WYSIWYGエディタ)を使って文書を作成し、コンバータを使って、所望の論理構造を得るという方法が広く用いられている。この場合、テキストエディタやWYSIWYGエディタでの文書作成は一定の規則に沿って行う必要がある。

【0009】作成規則は、大きく分けて、構造を抽出するために予め定められたパターンに合うよう文章を作成する方法と、要素として扱う部分に特定のスタイル(体裁を指定するための情報)を指定する方法の2つが用いられる。

【0010】パターンを用いる場合には、例えば、要素と要素の間には空行(連続する改行)を入れる、特定の要素はインデントをつけて表現する、正規表現などにより予め定められたパターンに合うよう項目に番号づける、正規表現などにより予め定められた文字列を用いる、などの方法が取られている。

【0011】他方、スタイルを用いる場合には、要素として用いる段落、文、語句などに、予め定められたスタイルを指定する。ところで、文書の検索を行うには、文書がもつテキストを検索条件に用いる全文検索の手法が用いられることが多い。全文検索では、文書がもつ語句に関する条件を、AND(かつ)、OR(または)、NOT(否定)で結合したブーリアン検索が一般的である。単純なブーリアン検索では、再現率(recall)はよいが、精度(precision)が低くなる傾向がある。すなわち、検索結果に、ユーザが期待していない文書が数多く含まれることがよくあるという問題がある。

【0012】そこで、構造化文書を対象とする文書検索においては、論理構造を用いることにより、検索精度をあげる方法が広く用いられている。例えば、章見出しに「文書処理」という文字列をもつ章にあり、図見出しに「データベース」という図を検索することにより、文書データベースや、データベース中のデータをもとに生成した文書などといった、文書処理の文脈でのデータベースに関する図を検索することができる。

【0013】これらの手法を用いている従来技術には、特開平4-217073号公報(文書蓄積システムにおける文書検索装置)がある。この文書検索装置では、文書の論理構造におけるノードの親子関係、ノードの内容および属性を用いて、検索対象を指定できる。これにより、論理構造が正しく作成されていれば、構造を利用することにより、再現率を低下させることなく精度を上げることができる。

【0014】

【発明が解決しようとする課題】しかし、従来の構造化文書を対象とする文書検索装置では、論理構造が正しく作成されていない場合には、精度が低下することはないが再現率が低下してしまうという問題点があった。

【0015】即ち、文書はユーザが作成するものであるから、論理構造が常に正しく作成されているとは限らない。とくに、WYSIWYGエディタを用いて文書を作成している場合には、所定のパターンにマッチしないように記述されたり、画面表示が同一になるか類似した表示になる、論理構造の異なる複数の文書要素が混同されたりする。この結果、文書型の制約は満たしているが、本来もつべき論理構造とは異なる論理構造をもつ文書が作成される。

【0016】以上の理由により、構造化文書を対象とする従来の文書検索装置では、文書型によって定められた要素型の情報を用いて多数の文書を対象として検索すると再現率が低下する。このことは、先に例示した従来技術以外の技術も含めた、既存の全ての文書検索装置にあてはまる。

【0017】本発明はこのような点に鑑みてなされたものであり、論理構造の異なる複数の文書に対する検索においても、高い再現率を維持した構造化文書検索装置を提供することを目的とする。

【0018】

【課題を解決するための手段】本発明では上記課題を解決するために、構造化文書を検索対象とする構造化文書検索装置において、文書構造中のノードのタイプ、ノードの内容、ノードの属性、ノード間の構造上の関係によって記述された検索式が入力されると、前記検索式に示された検索条件を段階的に緩やかな条件に書き換えた条件緩和検索式を生成する検索式生成手段と、前記条件緩和検索式を生成するのに行った書き換えの内容に応じて、前記条件緩和検索式による検索結果の確からしさを示す確度を計算する確度計算手段と、入力された前記検索式及び前記条件緩和検索式に基づいて検索を行う検索実行手段と、前記検索実行手段による検索結果を、確度の高い順に並べて併合する検索結果併合手段と、を有することを特徴とする構造化文書検索装置が提供される。

【0019】このような構造化文書検索装置によれば、検索式が入力されると、検索式生成手段によって、前記検索式に示された検索条件を段階的に緩やかな条件に書き換え、条件緩和検索式が生成される。すると、確度計算手段により、各条件緩和検索式を生成するのに行った書き換えの内容に応じて、条件緩和検索式による検索結果の確からしさを示す確度が計算されるとともに、検索実行手段により、入力された検索式及び条件緩和検索式に基づいて検索が行われる。検索結果は、検索結果併合手段により、確度の高い順に並べて併合される。

【0020】また、構造化文書を検索対象とする構造化

文書検索装置において、文書構造中のノードのタイプ、ノードの内容、ノードの属性、ノード間の構造上の関係によって記述された検索式が入力されると、前記検索式内の各検索条件を個別の部分式とし、前記部分式毎に条件を満たした文書部品の取り出しを行う部分式評価手段と、前記部分式評価手段で取り出された文書部品が条件を満たす部分式に応じて、各文書部品の確からしさの度合いを示す確度を計算する確度計算手段と、確度の高い文書部品の順に検索結果を出力する検索結果出力手段と、を有することを特徴とする文書検索装置が提供される。

【0021】このような構造化文書検索装置によれば、検索式が入力されると、部分式評価手段によって、検索式内の各検索条件を個別の部分式とし、部分式毎に条件を満たした文書部品の取り出しが行われる。すると、確度計算手段により、部分式評価手段で取り出された文書部品が条件を満たす部分式に応じて、各文書部品の確からしさの度合いを示す確度が計算される。そして、検索結果出力手段により、確度の高い文書部品の順に検索結果が出力される。

【0022】

【発明の実施の形態】以下、本発明の実施の形態を図面を参照して説明する。図1は、本発明の原理構成図である。まず、文書構造中のノードのタイプ、ノードの内容、ノードの属性、ノード間の構造上の関係によって記述された検索式1が入力されると、検索式生成手段2が、検索式に示された検索条件を段階的に緩やかな条件に書き換え、条件緩和検索式4を生成する。例えば、構造条件の「AND」を「OR」に書き換えれば、検索条件が緩和された条件緩和検索式となる。このような条件緩和検索式4が生成されると、確度計算手段3は、各条件緩和検索式4を生成するのに行った書き換えの内容に応じて、条件緩和検索式4による検索結果の確からしさを示す確度を計算する。

【0023】検索実行手段5は、入力された検索式1と検索式生成手段2により生成された条件緩和検索式4のそれぞれにより、文書保持手段6内の構造化文書を対象として検索を実行する。検索結果併合手段7は、検索実行手段5による検索結果を、確度の高い順に並べて併合する。

【0024】これにより、正しく論理構造が作成されていない文書も検索することができる。単純に検索結果を見た場合にはもとの検索式で検索したときに比して適合率が低下するが、検索結果はもとの検索式に正確に一致するものとあまり一致しないものとが乱雑に並ぶことなく、もとの検索式に照らして確からしい順に出力されるので、利用者は確からしいものから順次検索結果を吟味することができる。

【0025】図2は、構造化文書検索装置の第1の実施の形態のブロック図である。この実施の形態は、以下の

ような構成要素からなる。なお、この実施の形態では、図1の説明で用いた「確度」は、「ペナルティ」で表している。このペナルティは、値が小さいほど検索結果の確からしさが大きくなる。

【0026】検索式入力手段11は、利用者に対して、検索式の作成及び入力機能を提供している。また、この検索式入力手段11は、他のプログラムで作成した検索式を入力することもできる。ペナルティ割当手段12は、利用者に対して、ペナルティを計算する基準となる値を検索式の書き換え規則毎に割り当てるための機能を提供している。ペナルティ計算手段13は、ペナルティ割当手段12により入力されたペナルティの値に従い、各検索式のペナルティの合計値を求める。ペナルティ指定手段14は、利用者に対して、ペナルティの上限値の入力機能を提供する。

【0027】検索式生成手段15は、検索式入力手段11により入力された検索式をもとに、検索式書き換え規則にしたがって検索条件の緩やかな検索式を生成する。検索式保持手段16は、検索式入力手段11により入力された検索式と検索式生成手段15で生成された検索式とを、ペナルティ計算手段13により計算されたペナルティの値を付加して保持する。

【0028】文書保持手段17には、論理構造の異なる複数の文書が格納されている。検索実行手段18は、文書保持手段17に保持された文書を対象とし、検索式保持手段16に格納されている検索式による検索を行う。検索結果保持手段19は、検索式実行手段18が行った検索結果を格納する。この際、各検索結果には、検索式に設定されたペナルティの値が付加されている。検索結果併合手段20は、検索結果保持手段19に格納されている検索結果を、付加されたペナルティの値の小さい順に並べ替える。検索結果出力手段21は、検索結果併合手段20で並べ替えられた順に、検索結果を表示装置の画面上に表示する。

【0029】このような構成の構造化文書検索装置において検索を実行するには、予めペナルティ割当手段12を用いて、検索式の書き換え規則に対するペナルティを設定する。設定するペナルティは、検索式の書き換えによって、検索の条件が緩やかになる度合いが大きいほど大きな値とする。入力されたペナルティは、ペナルティ計算手段13に渡され、そこで保持される。また、利用者は、ペナルティの上限値を、ペナルティ指定手段14を用いて入力することもできる。なお、ペナルティの上限値は必須の設定項目ではない。

【0030】以上の設定を行った後、文書検索をしようとする利用者は、検索式入力手段11により、任意の検索式を検索式生成手段15に入力する。すると、検索式生成手段15は、入力された検索式から、検索条件を段階的に緩やかにした検索式を多数生成する。検索式生成手段15で生成された検索式は、検索式保持手段16で

一時的に保持される。なお、ペナルティの上限値が設定されている場合には、ペナルティが上限値を超えた検索式が生成されることはない。

【0031】検索式保持手段16に保持された検索式は、検索式実行手段18により順に取り出され、文書保持手段17に格納されている構造化文書を対象として検索が行われる。そして、各検索式を評価した結果得られる検索結果は、検索式に付加されていたペナルティの値とともに検索結果保持手段19に保持される。

【0032】すべての検索式が評価されたのち、ペナルティとともに検索結果保持手段19に保持された検索結果は、検索結果併合手段20により、ペナルティの値の小さい順に並べ替えて併合される。併合された検索結果は、検索結果出力手段21により、順次出力される。

【0033】次に、第1の実施の形態の内容を具体的に説明する。まず、書き換えの規則に割り当てるペナルティの例を示す。図3は、書き換え規則に割り当てられたペナルティの例を示す図である。このペナルティ13aは、ペナルティ割当手段12で入力され、ペナルティ計算手段13が保持するものである。この例では、書き換え種別が「親子→祖孫」の場合にはペナルティは「10」である。書き換え種別が「タイプの無視」の場合にはペナルティは「30」である。書き換え種別が「子孫のAND→OR」の場合にはペナルティは「100」である。書き換え種別が「テキストに関する条件のAND→OR」の場合にはペナルティは「200」である。書き換え種別が「ブーリアン検索」の場合にはペナルティは「500」である。

【0034】このようなペナルティが設定された状態で、利用者は、検索式入力手段11により検索式を入力する。例えば、以下のような検索式を入力する。図4は、検索式の例を示す図である。図に示すように、利用者が検索式入力手段11を用いて入力した検索式30は、内部的には有向グラフで表現される。

【0035】AND以外のラベルをもつノード31、33～35は、文書のある要素自身に関する条件(局所条件)と、そのノードにもっとも近い、AND以外のラベルをもつノードの局所条件で指定された要素との構造上の関係を示す条件(構造条件)を表現する。この例で示した局所条件は、取出対象であるか否か(真、偽)の条件、そのノードのタイプの種別に関する条件、及びそのノードのテキストに関する条件である。また、構造条件は、上位のノードとの関係が親子であるか、祖孫であるかにより指定されている。

【0036】ラベルがANDであるノード32は、構造に関する条件の連言を示す。なお、検索式入力手段11により入力された検索式のペナルティは、常に「0」である。

【0037】以後、図4の有向グラフのノードおよびノード内で指定された条件を、もとの検索式の部分式と呼

ぶ。図4のような検索式30が入力されたら、検索処理の実行に先だって、検索式の書き換えが行われる。

【0038】図5は、検索式の書き換え処理のフローチャートである。以下、図5の手順にしたがい、検索式の書き換え処理について説明する。

【S1】ペナルティ計算手段13は、ペナルティの値を初期化する。即ち、値を「0」にする。

【0039】以下のステップS2ないしステップS9は、検索式の内部表現の部分式に対する繰り返し処理である。

【S2】検索式生成手段15は、現在与えられている検索式において、適用可能な書き換え規則が存在するか否か検査する。適用可能な書き換え規則が存在すればステップS3へ、そうでなければ実行を終了する。

【S3】検索式生成手段15は、適用可能な書き換え規則をひとつ選択する。

【S4】検索式生成手段15は、ステップS3で選択された書き換え規則がタイプの無視であるか否か検査する。書き換え規則が「タイプの無視」であればステップS5へ、そうでなければステップS6へ行く。

【S5】検索式生成手段15は、取出対象以外のノード(取出対象が「偽」)のタイプに関する条件を、任意型を示す「ANY」に書き換える。すなわち、タイプに関する条件を無効化する。この処理の後、ステップS7へ行く。

【S6】検索式生成手段15は、ステップS3で選択された書き換え規則を適用し、検索式を書き換える。書き換え規則が「親子→祖孫」、「子孫のAND→OR」、若しくは「テキストに関する条件のAND→OR」のいずれかであれば、その書き換え規則の通りに検索式を書き換えを行う。また、書き換え規則が「ブーリアン検索」の場合には、検索式全体をブーリアン検索に変更する。その変更の手続については、後述する(図10に示す)。

【S7】ペナルティ計算手段13は、現在処理中の検索式に対し、ステップS3で選択された書き換え規則に対応するペナルティを加算する。

【S8】検索式生成手段15は、ペナルティが基準値を超えたか否かを検査する。ペナルティが基準値を超えていれば実行を終了し、そうでなければステップS9へ行く。

【S9】ステップS5あるいはステップS6の検索式を書き換えて得られた検索式を、検索式保持手段16が保存し、ステップS2へ行く。

【0040】以上のような処理が行われることにより、図4に示した検索式から複数の新たな検索式が生成される。そして、各検索式には、書き換えの内容に応じてペナルティが付加される。図4の検索式30を書き換えた例を図6ないし図9に示す。

【0041】図6は、書き換え後の検索式の第1の例を

示す図である。この検索式30aは、図4の検索式30に対し、親子関係を祖孫関係に変更する書き換え規則を3箇所に適用したものである。この場合のペナルティは $10 \times 3 = 30$ となる。

【0042】図7は、書き換え後の検索式の第2の例を示す図である。この検索式30bは、図6の検索式30aに対し、タイプを無視する書き換え規則を3箇所に適用したものである。この場合のペナルティは(図6の検索式30aのペナルティ) $+ 30 \times 3 = 30 + 90 = 120$ である。

【0043】図8は、書き換え後の検索式の第3の例を示す図である。この検索式30cは、図7の検索式30bに対し、子孫に関するANDの条件をORに変更する書き換え規則を1箇所に適用したものである。この場合のペナルティは、(図7の検索式30bのペナルティ) $+ 100 = 120 + 100 = 220$ である。

【0044】図9は、書き換え後の検索式の第4の例を示す図である。この検索式30dは、図8の検索式30cに対し、テキストに関する条件のANDをORに変更する書き換え規則を1箇所に適用したものである。この場合のペナルティは、(図8の検索式30cのペナルティ) $+ 200 = 220 + 200 = 420$ である。

【0045】図6から図8は、木構造の検索式から木構造の検索式への書き換えを行った場合の例であり、これらの書き換えは、各ノード内の検索条件を、書き換え規則に従って書き換えればよい。一方、ブーリアン検索への書き換えを行うには、次のような手続を実行する必要がある。

【0046】図10は、検索式全体をブーリアン検索へ書き換える手続きのフローチャートである。この手続は検索式生成手段15の行う処理であり、この手続きの入力木構造で表現された検索式のノードで、出力はそのノードを頂点とする検索式の部分木を交換して得られたブーリアン検索式である。この手続きでは、出力される検索式は文字列で表わされる。

【S11】ノードが要素に対する条件か否かを判定する。要素に関する条件であればステップS12へ、そうでなければステップS13へ行く。

【S12】ノードに指定されたテキスト内容に関する条件を検索式とし、ステップS14へ行く。

【S13】検索式を空とし、ステップS14へ行く。

【S14】ノードが子ノードをもつか否かを判定する。子ノードが存在すればステップS15へ、存在しなければステップS23へ行く。

【0047】以下のステップS15ないしステップS22は、ノードの子ノードに対する繰り返し処理である。

【S15】交換処理を施していない子ノードがあるか否かを判定する。未処理の子ノードがあればステップS16へ、そうでなければステップS23へ行く。

【S16】未処理の子ノードを1つ選択する。

【S17】ステップS16で選択したノードを引数として、ブーリアン検索への書き換え手続きを再帰的に呼び出す。

【S18】検索式が空であるか否かを判定する。検索式が空であればステップS19へ、そうでなければステップS20へ行く。

【S19】ステップS17の手続き呼出しの結果得られた検索式(文字列)を検索式とし、ステップS15へ行く。

10 【S20】この手続きの引数として与えられたノードがORノードであるか否かを判定する。ORノードであればステップS22へ、そうでなければステップS21へ行く。

【S21】検索式と、ステップS17の手続き呼出しの結果得られた検索式とをANDで連結し、ステップS15へ行く。

【S22】検索式と、ステップS17の手続き呼出しの結果得られた検索式をORで連結し、ステップS15へ行く。

20 【S23】検索式を戻り値として、手続きの実行を終了する。

【0048】このような処理により、木構造の検索式からブーリアン検索式を求めることができる。以下に、図4の検索式を直接ブーリアン検索式へ書き換えた例を示す。

【0049】

【数1】(SQL AND conference) AND "Yuri Rubinski" AND Publishing

このブーリアン検索式のペナルティは、(図4の検索式30のペナルティ) $+ 500 = 0 + 500 = 500$ である。

【0050】以上のような、複数の木構造の検索式とブーリアン検索式とのそれぞれに基づいて、検索が行われると、それぞれの検索結果に対して検索式のペナルティが付加される。そして、全ての検索結果がペナルティの低い順に並べられる。この際、1つの文書が複数の検索式により検出された場合には、その検索結果には、値の小さい方のペナルティが採用される。並べられた検索結果は、表示装置の画面に表示される。

40 【0051】図11は、第1の実施の形態による検索結果の表示例を示す図である。同図では、検索結果の表示画面21aの中に、検索結果として得られた文書が表示されている。これらの文書は、ペナルティ、文書名、著者、作成日とともに表示されている。画面中にはスクロールバーが設けられており、このスクロールバーを操作することにより、表示させる文書をスクロールさせ、ペナルティの小さい検索結果を順次画面表示させることができる。ペナルティの上限値が定められていれば、その上限値を超える検索結果は存在しない。

50 【0052】なお、同図ではペナルティにより検索条件

の厳しさを表示しているが、検索条件が厳しいものほど高い値となるようなスコアを計算して表示してもよい。例えば、ペナルティ指定手段14によりペナルティの上限値を指定した場合には、その上限値からペナルティの値を引いたものをスコアとすることができる。

【0053】このようにして、入力した検索式から検索条件を段階的に緩めた検索式による検索結果を閲覧することができる。従って、正しい論理構造に基づいて作成した検索式を入力した場合でも、論理構造が正しく作成されていない文書を検出することができ、再現率が向上する。しかも、検索式の条件を緩める度合いが低いものを優先的に表示するため、検索結果の数が多くなっても利用者の閲覧が不便になることはない。

【0054】ところで、上記の第1の実施の形態は、入力された検索式の検索条件を書き換えることにより、段階的な検索条件の緩和を行ったものであるが、入力された検索式の部分式に基づいて文書部品（構造化文書の個々の要素）を取り出し、検索式の条件を満たしている度合いの高い文書部品を、検索結果として出力することもできる。そのような例を第2の実施の形態として以下に説明する。

【0055】図12は、構造化文書検索装置の第2の実施の形態のブロック図である。なお、この実施の形態では、検索結果の確からしさを「スコア」で表している。このスコアは、値が大きいほど検索結果の確からしさが大きいことを示す。検索式入力手段41は、利用者に対して、検索式の作成及び入力機能を提供している。検索式保持手段42は、検索式入力手段41により入力された検索式を保持する。

【0056】文書保持手段43には、論理構造の異なる複数の文書が格納されている。部分式評価手段44は、検索式保持手段42に格納された検索式の部分式を順次取り出し、その部分式により文書保持手段43に格納された文書を対象として評価を行う。即ち、取り出した部分式の示す条件を満たした文書部品を文書保持手段43から候補として取り出す。一方、スコア割当手段45は、スコアを計算する基準となる値を部分式の種類毎に割り当てるための機能を利用者に対して提供している。スコア計算手段46は、スコア割当手段45により入力されたスコアの値に従い、部分式評価手段44の取り出した候補のスコアを求める。

【0057】候補保持手段47は、部分式評価手段44が取得した候補に、スコアの値を付加して保持する。スコア指定手段48は、利用者に対して、スコアの下限値の入力機能を提供する。検索結果出力手段49は、スコア指定手段48により指定された下限値よりも大きいスコアの文書部品を検索結果として、スコアの大きいものから順に、表示装置の画面上に表示する。

【0058】なお、図中の部分式評価手段44、スコア計算手段46、及び候補保持手段47により、検索式評

価手段40を構成している。このような構造化文書検索装置において検索を実行するには、利用者は、予めスコア割当手段45を用いて、スコアを計算する基準となる値を部分式の種類毎に割り当てる。割り当てるスコアは、その検索条件により検索対象が絞られる度合いが大きいほど大きな値とする。入力されたスコアは、スコア計算手段46に渡され、そこで保持される。また、利用者は、スコアの下限値を、スコア指定手段48を用いて入力することもできる。なお、スコアの下限値は必須の設定項目ではない。

【0059】以上の設定を行った後、文書検索をしようとする利用者は、検索式入力手段41により、自己が作成したか若しくは所定のプログラムにより作成された検索式を、検索式保持手段42に入力する。検索式入力手段41で入力された検索式は、検索式保持手段42で保持される。検索式保持手段42に保持された検索式は、部分式評価手段44により、文書保持手段43に保持された文書を対象として、部分式ごとに評価される。部分式評価手段44で検索式の部分式を評価した結果得られた候補は、部分式毎にスコア計算手段46により算出されたスコアとともに候補保持手段47に保持される。部分式を評価していく段階で、候補となる文書部品が得られたとき、その部品がまだ候補保持手段47に存在していなければ、スコア計算手段46はその候補のスコアとして該部分式のスコアを割り当てる。得られた候補がすでに候補保持手段47に存在していれば、スコア計算手段46は該候補のスコアを該部分式のスコア分だけ増加させる。

【0060】すべての部分式が評価されたのち、スコアとともに候補保持手段47に保持された検索結果は、検索結果出力手段49によりスコア順に出力される。次に、第2の実施の形態の内容を具体的に説明する。まず、文書を評価する条件に割り当てるスコアの例を示す。

【0061】図13は、文書を評価する条件に割り当てられたスコアの例を示す図である。このスコア46aは、スコア割当手段45で入力され、スコア計算手段46が保持するものである。図の例では、条件が「タイプ」の場合にはスコアは「100」である。条件が「親子関係」の場合にはスコアは「100」である。条件が「祖孫関係」の場合にはスコアは「100」である。条件が「テキストに関する条件」の場合にはスコアは「100」である。条件が「タイプのAND」の場合にはスコアは「200」である。条件が「テキストに関する条件のAND」の場合にはスコアは「200」である。

【0062】このようなスコアが設定された状態で、利用者は、検索式入力手段41により検索式を入力する。例えば、図4のような検索式を入力する。すると、検索式評価手段40は、文書保持手段43内の文書に対して検索式の評価を行い、評価の高い文書を検索結果とす

る。

【0063】図14は、検索式を評価する手続きのフローチャートである。この手続きでは、検索条件を評価する段階で、検索結果の候補に対しスコアを与えていく。

【S31】部分式評価手段44は、検索式の内部表現のノードのうち、未処理のものが存在するか否かを検査する。未処理のノードが存在すればステップS32へ行く。そうでなければ、ステップS37に行く。

【S32】部分式評価手段44は、検索式の内部表現のノードのうち、未処理のものを1つ選択する。

【S33】部分式評価手段44は、ステップS32で選択されたノードがもつ検索条件のうち、未処理のものが存在するか否かを検査する。未処理の条件が存在すればステップS34へ、そうでなければステップS31へ行く。

【S34】部分式評価手段44は、ステップS32で選択されたノードがもつ検索条件のうち、未処理のものを1つ選択する。

【S35】部分式評価手段44は、ステップS34で選択された検索条件を評価する。

【S36】スコア計算手段46は、ステップS35で検索条件を評価した結果として得られた各候補に対し、ステップS35で評価した検索条件に対応するスコアを加算する。その後、ステップS33へ行く。

【S37】部分式評価手段44は、構造条件のうち、未評価のものが存在するか否かを検査する。未評価の構造条件が存在すればステップS38へ、そうでなければステップS41へ行く。

【S38】部分式評価手段44は、構造条件のうち、未評価のものを1つ選択する。

【S39】部分式評価手段44は、ステップS38で選択された構造条件を評価する。

【S40】スコア計算手段46は、ステップS39の評価結果得られた各候補に対し、ステップS39で評価した条件に対応するスコアを加算する。

【S41】候補保持手段47は、取出対象でないものを候補から除外して、残った候補を保持する。即ち、取出対象が「真」である候補のみを保持する。

【0064】以上のような処理が行われることにより、図4に示した検索式から検索結果が得られる。そして、各検索式には、スコアが付加される。なお、図4の検索式で取出対象となるのは、タイプが「Section」のノード31のみであるため、取り出される検索結果は、タイプが「Section」の文書部品だけである。図15ないし図17に、図4の検索式30による検索結果をスコアとともに示す。

【0065】図15は、検索結果の第1の例を示す図である。この文書部品50は、各ノード51～54の関係と内容とが、図4の検索式30の条件を全て満たしている。そのため、スコアは最大の値(=1500)とな

る。

【0066】図16は、検索結果の第2の例を示す図である。この文書部品50aは、3つのノード51、52、54に関する条件のみを満たしている。そのため、スコアは図15の例よりも低くなり、スコア=1000である。

【0067】図17は、検索結果の第3の例を示す図である。この文書部品50bは、2つのノード51、52に関する条件のみを満たしている。そのため、スコアは図16の例よりもさらに低くなり、スコア=700である。

【0068】このように取り出された検索結果は、スコアの順に表示装置の画面に表示される。図18は、第2の実施の形態による検索結果の表示例を示す図である。同図では、検索結果の表示画面49aの中に、検索結果として得られた文書が表示されている。これらの文書は、スコア、文書名、著者、作成日とともに表示されている。この画面は、スクロールバーを操作することにより、画面中表示させる文書をスクロールさせ、スコアの小さい検索結果を順次画面表示させることができる。このとき、スコアの上限値が定められていれば、その上限値を超えた検索結果は表示されない。

【0069】なお、以上の2つの実施の形態以外に、次のような変形例も考えられる。上記の実施の形態では、スコアやペナルティは固定の値としたが、データベースの状態により可変にしてもよい。以下に、データベースの状態によりこれらの値を変更する方法の例を示す。

【0070】例えば、テキストに関する条件の場合、データベース中に頻出する語句ほどスコアあるいはペナルティを下げる。また、タイプに関する条件の場合、データベース中に頻出する語句ほどスコアあるいはペナルティを下げる。

【0071】このように、語句の出現頻度に応じてペナルティ若しくはスコアを計算するには、ペナルティ計算手段13(図2に示す)若しくはスコア計算手段46(図12に示す)が、テキストに関する条件が指定されると、テキストに関する条件の語句の頻度の度合いを求める手段と、その度合いに応じたペナルティ若しくはスコアを求める手段とを有していればよい。

【0072】また、必然的に親子関係あるいは祖孫関係が成立する場合、親子関係あるいは祖孫関係に関する条件に一致したときのスコアは0とする。この場合には、親子関係あるいは祖孫関係が必ず成立するため、ペナルティの値がいくらになっていても実質的には意味をなさない。親子関係あるいは祖孫関係が必然的に成立するかどうかは、文書型を見れば判定できる。あるタイプT1から、SEQまたはREPノードだけを辿って別のタイプT2に到達できるとき、T1が存在すればT2が必ず存在し(逆も真)、それらは親子関係にある。また、あるタイプT1から、SEQ、REP、またはタイプノー

ドだけを辿って別のタイプT2に到達できるとき、T1が存在すればT2が必ず存在し(逆も真)、それらは祖孫関係にある。

【0073】

【発明の効果】以上説明したように第1の発明では、入力された検索式に基づいて、検索条件を段階的に緩やかにした条件緩和検索式を生成し、条件の緩やかな検索をも行うようにしたため、正しく論理構造が作成されていない文書も検索することができ、再現率が向上する。しかも、検索結果はもとの検索式に照らして確からしい順

に出力されるので、利用者は確からしいものから順次検索結果を吟味することができる。

【0074】また、第2の発明では、入力された検索式の部分式により文書部品を取り出し、入力された検索式の条件を満たしている度合いが高いものから順に検索結果として出力するようにしたため、上記第1の発明と同様に、正しく論理構造が作成されていない文書も検索することができ、再現率が向上する。

【図面の簡単な説明】

【図1】本発明の原理構成図である。

【図2】構造化文書検索装置の第1の実施の形態のブロック図である。

【図3】書き換え規則に割り当てられたペナルティの例を示す図である。

【図4】検索式の例を示す図である。

【図5】検索式の書き換え処理のフローチャートである。

【図6】書き換え後の検索式の第1の例を示す図である。

【図7】書き換え後の検索式の第2の例を示す図であ

＊る。

【図8】書き換え後の検索式の第3の例を示す図である。

【図9】書き換え後の検索式の第4の例を示す図である。

【図10】検索式全体をブーリアン検索へ書き換える手続きのフローチャートである。

【図11】第1の実施の形態による検索結果の表示例を示す図である。

【図12】構造化文書検索装置の第2の実施の形態のブロック図である。

【図13】文書の評価する条件に割り当てられたスコアの例を示す図である。

【図14】検索式を評価する手続きのフローチャートである。

【図15】検索結果の第1の例を示す図である。

【図16】検索結果の第2の例を示す図である。

【図17】検索結果の第3の例を示す図である。

【図18】第2の実施の形態による検索結果の表示例を示す図である。

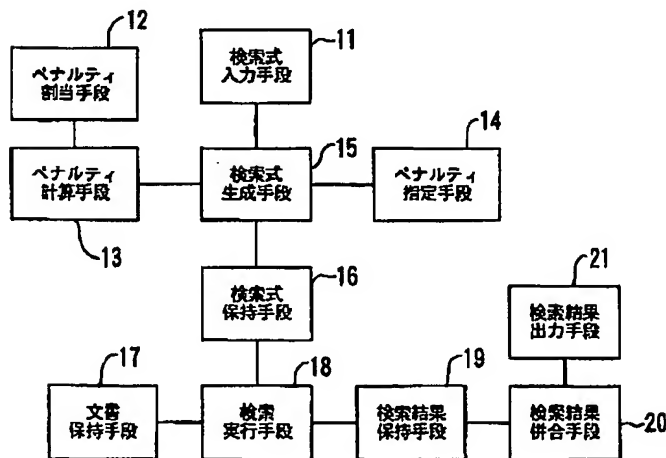
【図19】論理構造の例を示す図である。

【図20】文書型の例を示す図である。

【符号の説明】

- 1 検索式
- 2 検索式生成手段
- 3 確度計算手段
- 4 条件緩和検索式
- 5 検索実行手段
- 6 文書保持手段
- 7 検索結果併合手段

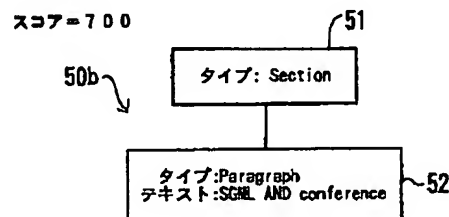
【図2】



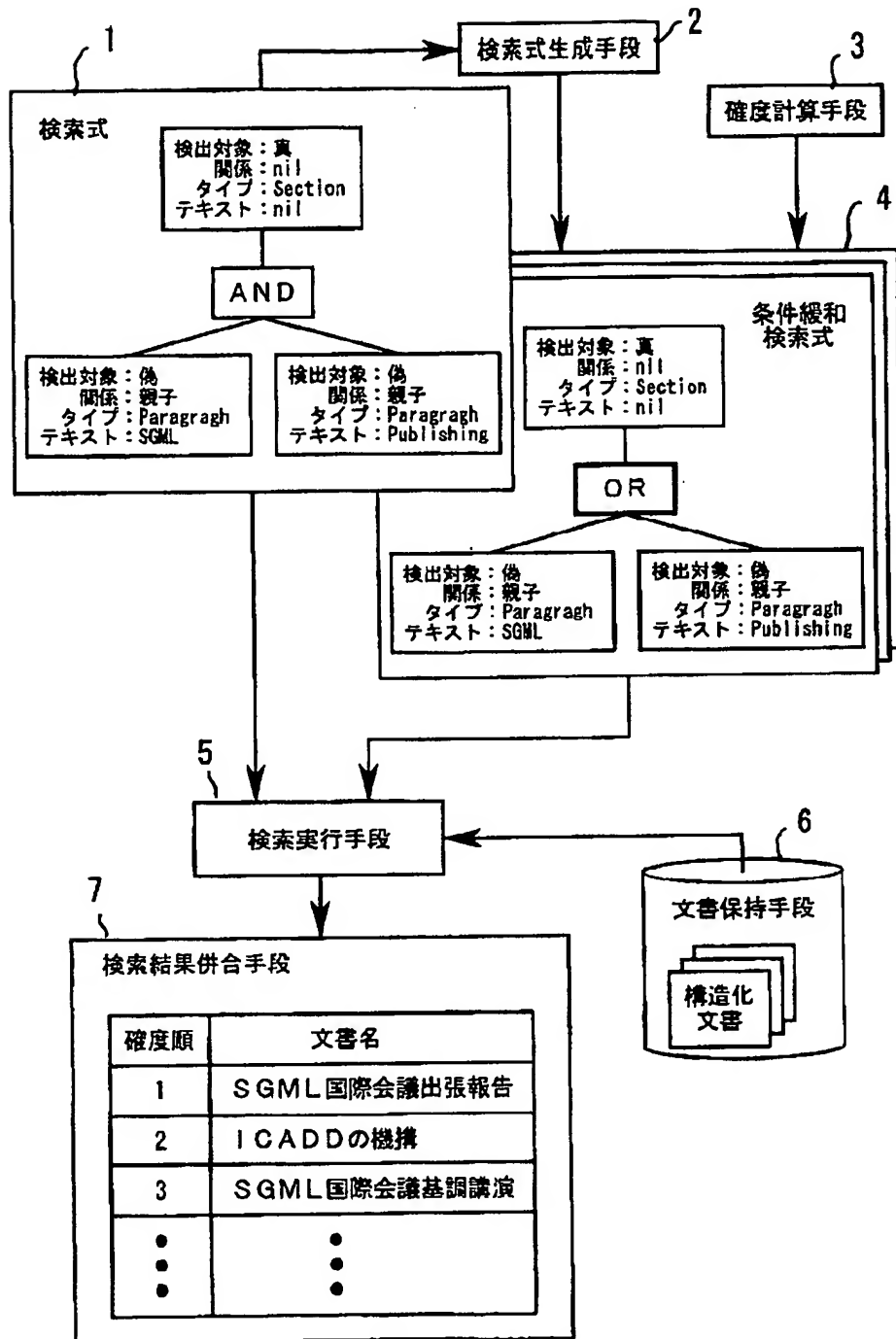
【図3】

書き換え種別	ペナルティ
親子→祖孫	10
タイプの無視	30
子孫のAND→OR	100
テキストに関する条件のAND→OR	200
ブーリアン検索	500

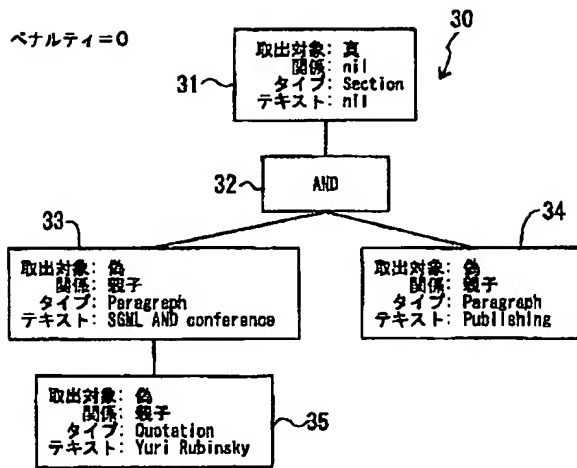
【図17】



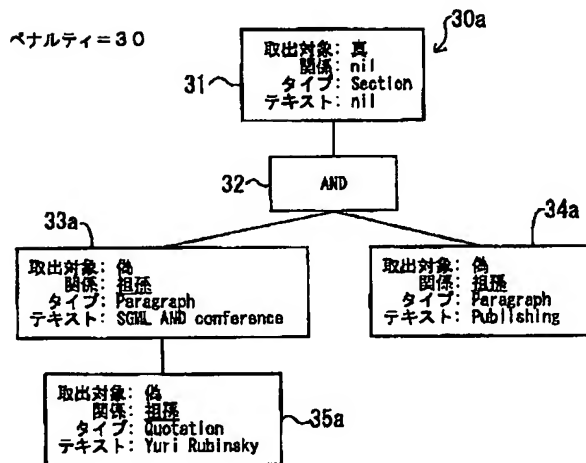
【図1】



【図4】



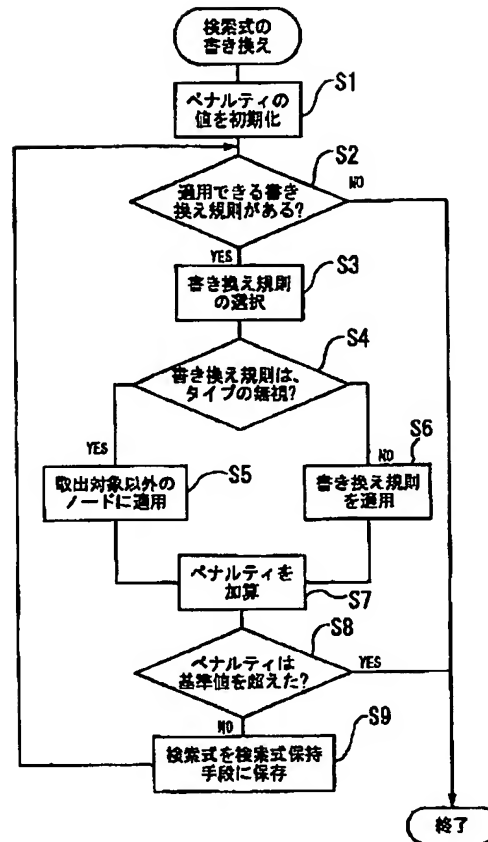
【図6】



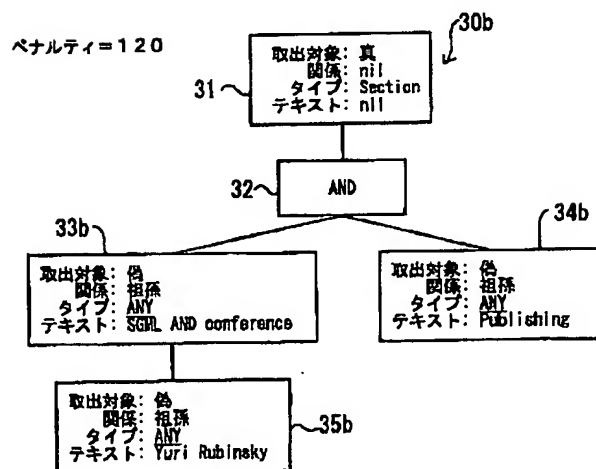
【図13】

条件	スコア
タイプ	100
親子関係	100
祖孫関係	100
テキストに関する条件	100
タイプのAND	200
テキストに関する条件のAND	200

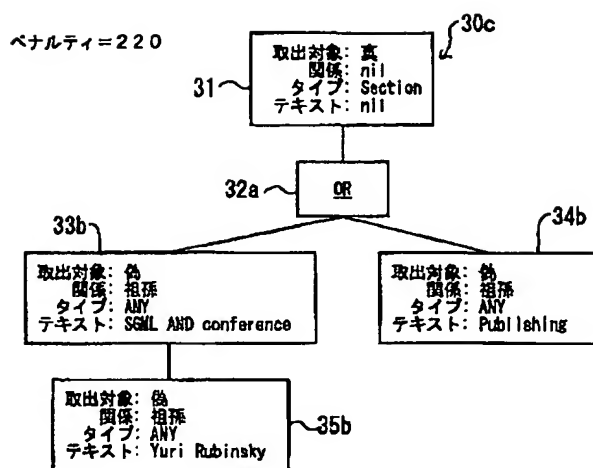
【図5】



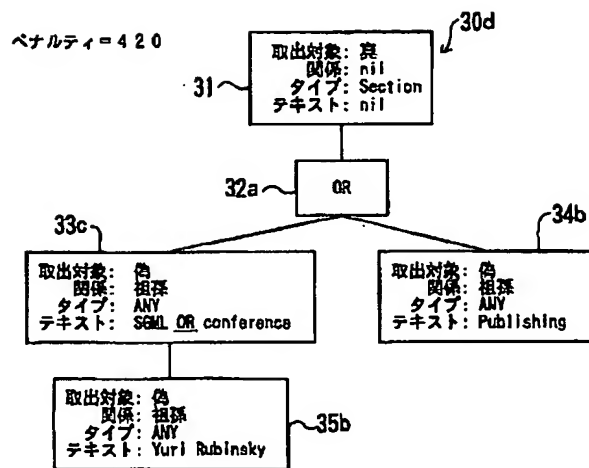
【図7】



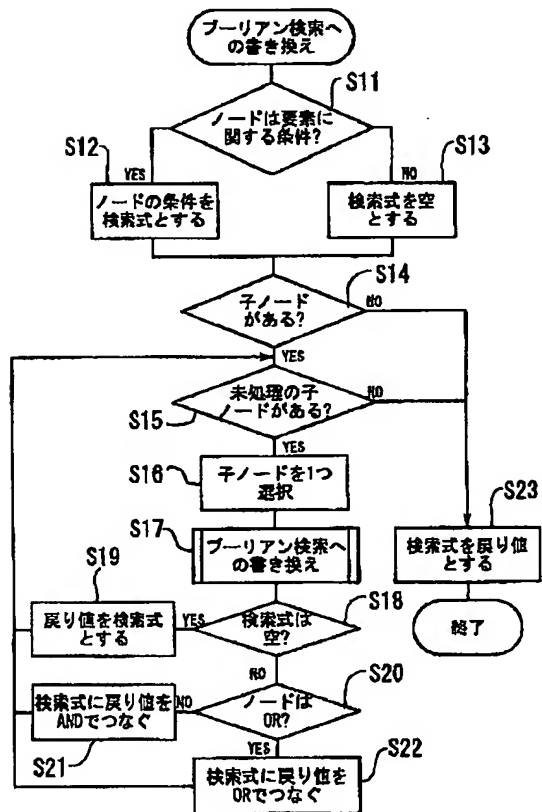
【図8】



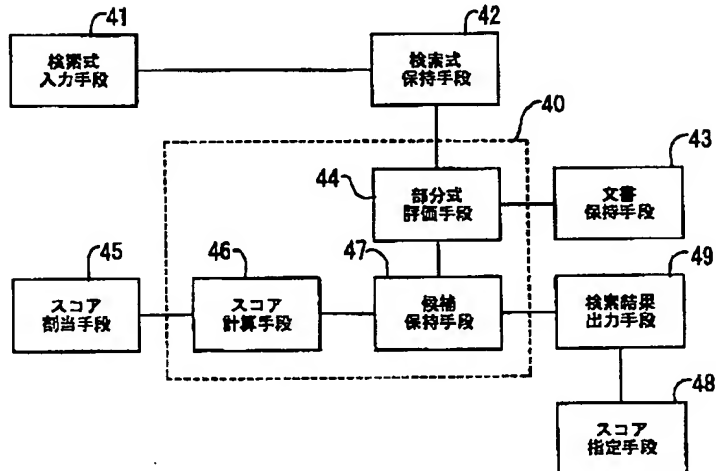
【図9】



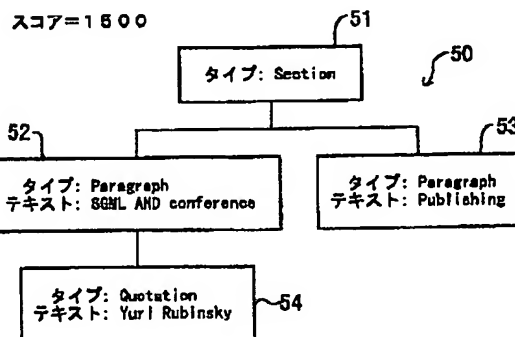
【図10】



【図12】



【図15】



【図11】

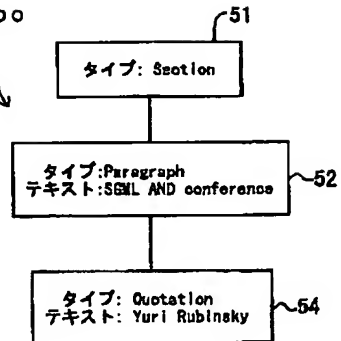
21a

検索結果				
番号	ペナルティ	文書名	著者	作成日
1	0	SGML国際会議出張報告	富士 太郎	96.11.27
2	30	ICADDの機構	富士 花子	96.1.20
3	120	SGML国際会議基調講演	富士 太郎	96.8.30
4	120	SGML国際会議基調講演の意図	富士 太郎	95.11.1
5	220	SGML勉強会資料2	富士 太郎	95.10.20
6	420	SGML勉強会資料1	富士 太郎	95.10.1

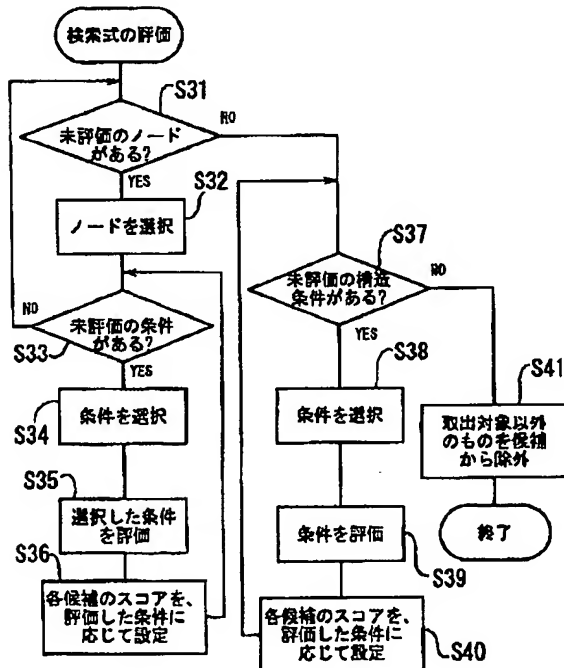
【図16】

スコア=1000

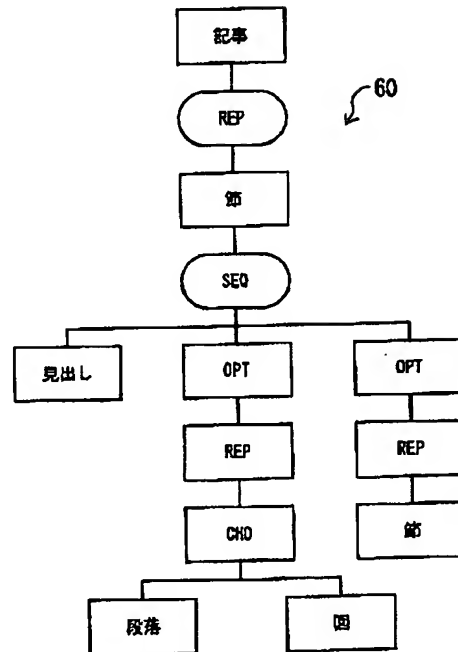
50a



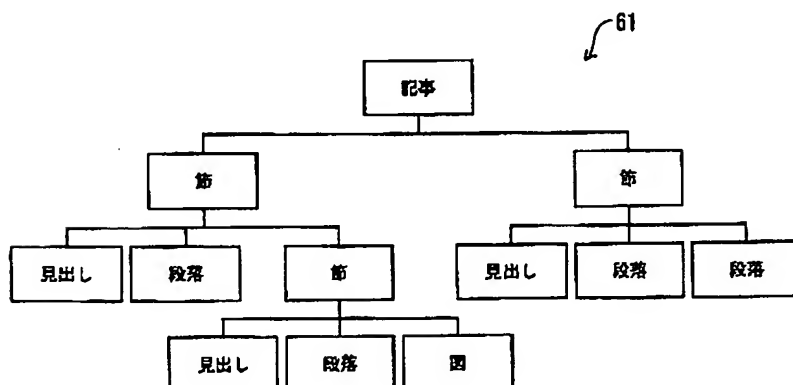
【図14】



【図20】



【図19】



【図18】

49a

検索結果				
番号	スコア	文書名	著者	作成日
1	1600	SGML国際会議出張報告	富士 太郎	96.11.27
2	1000	SGMLユーザの動向	富士 花子	96.1.20
3	700	SGMLの技術動向	富士 太郎	96.8.30
4	700	SGML勉強会資料2の補遺	富士 太郎	95.11.1
5	700	SGML勉強会資料2	富士 太郎	95.10.20
6	700	SGML勉強会資料1	富士 太郎	95.10.1

【公報種別】特許法第17条の2の規定による補正の掲載
 【部門区分】第6部門第3区分
 【発行日】平成14年1月25日(2002.1.25)

【公開番号】特開平10-198697
 【公開日】平成10年7月31日(1998.7.31)
 【年通号数】公開特許公報10-1987
 【出願番号】特願平9-4269
 【国際特許分類第7版】

G06F 17/30
 17/27
 17/21

【F I】

G06F 15/403 330 B
 15/20 550 E
 570 N
 15/40 370 A

【手続補正書】

【提出日】平成13年7月10日(2001.7.10)

【手続補正1】

【補正対象書類名】明細書

【補正対象項目名】発明の名称

【補正方法】変更

【補正内容】

【発明の名称】 構造化文書検索装置及び構造化文書検索方法

【手続補正2】

【補正対象書類名】明細書

【補正対象項目名】特許請求の範囲

【補正方法】変更

【補正内容】

【特許請求の範囲】

【請求項1】 構造化文書を検索対象とする構造化文書検索装置において、
 文書構造中のノードのタイプ、ノードの内容、ノードの属性、ノード間の構造上の関係によって記述された検索式が入力されると、前記検索式に示された検索条件を段階的に緩やかな条件に書き換えた条件緩和検索式を生成する検索式生成手段と、
 前記条件緩和検索式を生成するのに行った書き換えの内容に応じて、前記条件緩和検索式による検索結果の確からしさを示す確度を計算する確度計算手段と、
 入力された前記検索式及び前記条件緩和検索式に基づいて検索を行う検索実行手段と、
 前記検索実行手段による検索結果を、確度の高い順に並べて併合する検索結果併合手段と、
 を有することを特徴とする構造化文書検索装置。

【請求項2】 前記検索式生成手段は、予め定められた

限界値よりも高い確度の条件緩和検索式のみを生成することを特徴とする請求項1の構造化文書検索装置。

【請求項3】 書き換え規則ごとに基準確度を割り当てる確度割当手段をさらに有し、
 前記確度計算手段は、前記確度割当手段が割り当てた基準確度に基づいて、前記条件緩和検索式の確度を計算することを特徴とする請求項1記載の文書検索装置。

【請求項4】 構造化文書を検索対象とする構造化文書検索装置において、文書構造中のノードのタイプ、ノードの内容、ノードの属性、ノード間の構造上の関係によって記述された検索式が入力されると、前記検索式内の各検索条件を個別の部分式とし、部分式毎にそれぞれの条件を満たした文書部品の取り出しを行う部分式評価手段と、
 前記部分式評価手段で取り出された文書部品が、どのような部分式の条件を満たしているかに基づいて、各文書部品の確からしさの度合いを示す確度を計算する確度計算手段と、
 確度の高い文書部品の順に検索結果を出力する検索結果出力手段と、
 を有することを特徴とする文書検索装置。

【請求項5】 前記検索結果出力手段は、予め定められた限界値よりも高い確度の文書部品のみを検索結果とすることを特徴とする請求項4記載の構造化文書検索装置。

【請求項6】 部分式の種類ごとに基準確度を割り当てる確度割当手段をさらに有し、
 前記確度計算手段は、前記確度割当手段が割り当てた基準確度に基づいて、各文書部品の確度を計算することを特徴とする請求項4記載の構造化文書検索装置。

【請求項7】 構造化文書を検索対象とする構造化文書

検索方法において、
文書構造中のノードのタイプ、ノードの内容、ノードの
属性、ノード間の構造上の関係によって記述された検索
式の入力を受けて、前記検索式に示された検索条件を段
階的に緩やかな条件に書き換えた条件緩和検索式を生成
するステップと、
前記条件緩和検索式を生成するのに行った書き換えの内
容に応じて、前記条件緩和検索式による検索結果の確か
らしさを示す確度を計算するステップと、
入力された前記検索式及び前記条件緩和検索式に基づい
て検索するステップと、
前記検索の検索結果を確度の高い順に並べて併合するス
テップと、
を有することを特徴とする構造化文書検索方法。

【手続補正3】

【補正対象書類名】明細書

【補正対象項目名】0001

【補正方法】変更

【補正内容】

【0001】

【発明の属する技術分野】本発明は構造化文書を対象とした文書検索を行う構造化文書検索装置及び構造化文書検索方法に関し、特に複数の文書型から生成された文書を検索対象とする構造化文書検索装置及び構造化文書検索方法に関する。

【手続補正4】

【補正対象書類名】明細書

【補正対象項目名】0017

【補正方法】変更

【補正内容】

【0017】本発明はこのような点に鑑みてなされたものであり、論理構造の異なる複数の文書に対する検索においても、高い再現率を維持した構造化文書検索装置及び構造化文書検索方法を提供することを目的とする。

From: <inta2007@agip.com>
Date: 2/26/2007 11:40:13 AM
Subject: INTA 129th Annual Meeting, Chicago, April 28th - May 2nd, 2007

Dear Valued Clients:

We are pleased to inform you that Abu-Ghazaleh Intellectual Property (AGIP) will be participating in the INTA 129th Annual Meeting in Chicago - USA, April 28th - May 2nd, 2007.

AGIP's delegation will be staying at the Sheraton Chicago Hotel & Towers:

Luay T. Abu-Ghazaleh	Regional Office
Nabil Salame	Montreal Liaison Office
Charles Shaban	Regional Office
Mahmoud Lattouf	Regional Office
Suha Mahsiri	Regional Office
Christiane Bou Khater	Lebanon Office
Reham Mezieni	Egypt Office
Motasem Abu-Ghazaleh	UAE Office
Samer Pharaon	Saudi Arabia Office
Khalid Al-Khalidi	India Office
Aamir Khan	Pakistan Office
Khaled Battash	Morocco Office
Dima Naber	European Union Liaison Office
Afaf Shashaa	Turkey Office
Ridab Abu- Taleb	China Office

We will be hosting our meetings in rooms Parlor B, Parlor D, Parlor E, and Parlor F, at Level 3 in the Sheraton Chicago Hotel & Towers on April 30th, May 1st, and May 2nd, 09:00AM - 6:00PM.

Associates and clients who wish to have prior appointments with members of our delegation may please contact us at INTA2007@agip.com, by fax at +962 6 5100 901, or by calling us at +962 6 5100 900 ext. 1427.

Our delegation will be pleased to discuss any matters relating to any of our 60 offices.

Looking forward to seeing you in Chicago.

Very Truly Yours,

Abu-Ghazaleh Intellectual Property (AGIP)

Abu-Ghazaleh Intellectual Property (AGIP) provides an extensive range of intellectual property services from over 60 offices and through over 180 correspondents worldwide. For more information, please visit www.agip.com.

Our sister company, Abu-Ghazaleh Legal (ABLE), is now operating in all countries in the region. For more information, please visit

www.tag-legal.com

"We try harder to stay first"